Comparative Analysis of Classification Algorithms on Breast Cancer Dataset

Jyoti Negi[1], Dr. K.L. Bansal[2]

[1]*Department of Computer Science, HPU Shimla, India*
[2]*Department of Computer Science, HPU Shimla, India*

Abstract – *Breast cancer is one of the leading diseases among women nowadays. The abnormal growth of cells in breast tissue causes cancer in women. Early detection of breast cancer and breast cancer recurrence can save the patient. Classification algorithms are used to classify whether cancer is recurrent or non-recurrent. This paper presents the comparative analysis of various classification algorithms viz. K-Nearest Neighbors (KNN), Naïve Bayes and Random Forest. When these three classification algorithms are applied to the breast cancer dataset in the WEKA data mining tool for classifying recurrent and non-recurrent breast cancer, it has been observed that KNN has the highest True Positive rate (0.738) and Random forest has the lowest True Positive rate (0.696). It is observed that the performance of the K-Nearest Neighbors classifier algorithm is highest and the second most accurate classifier is Naïve Bayes with correctly classified instances of 71.6783%.*

Keywords: *classification algorithms, K-Nearest Neighbors, Naïve Bayes, Random Forest, breast cancer, WEKA*

I. INTRODUCTION

Data mining is the Extraction of useful information or finding the hidden pattern from a large amount of data. In the health care sector, huge amount of data is generated on the daily basis. The Bulk of patient records, reports and other informative data is generated. The major role of Data mining is to find the relevant information from a large number of databases. Various data mining and machine learning techniques were used for the extraction of relevant information from the large dataset. Popular data mining and machine learning techniques are classifications, clustering, regression, and association. For implementation WEKA tool is used in this research. Breast cancer recurrence and non-recurrence dataset are used for this research purpose. Breast cancer is the second most common leading disease among women and is a leading cause of death. Breast cancer can be recurrent or non-recurrent.

In this paper, for predicting breast cancer recurrence a model is developed by using data mining algorithms. The paper is divided into various sections, section II is related to literature survey, section III is related to proposed work, section IV is related to methodology and section V and section VI represents the experiment result and the conclusion.

II. LITERATURE REVIEW

Nidhi Sharma, K. L. Bansal [1] conducted a case study on classification algorithms and compared various data mining tools. The author used three different datasets which are having different numbers of attributes and different numbers of instances. They used classification accuracy for evaluating the tool. **Shelly Gupta, Dharminder Kumar, and Anand Sharma [2]** studied the performance analysis of various data mining classifiers KNN, NB, SVM, CART, DT, MLP, etc., and used three machine learning tools WEKA, Tanagra, Clementine. The author used four UCI machine learning repository datasets and 10 fold cross-validation used for the experiment. **Sujata Joshi and Mydhili K. Nair [3]** presented the classification-based data mining techniques applied to healthcare data. The author focuses on the prediction of heart disease using three classification techniques i.e. Decision Trees, Naïve Bayes, and K-Nearest Neighbors. The result shows that KNN has the highest accuracy. But when used for prediction the Decision Tree performs well when compared to the other two methods for the given heart disease dataset. **Nour A. AbouElNadar and Amani A. Saad [4]** compared the performance of the ensemble method i.e. Bagging, Voting, and Random Forest, and used four classification techniques i.e. KNN, DT, NB, SVM. The recurrence and non-recurrence dataset of the UCI machine learning repository of Breast cancer was used and the experiment was done on the WEKA tool. The result shows that the voting ensemble method gives the best result with 89.9% accuracy. **Vikas Chaurasia and Saurabh Pal [5]** experimented on the WBC dataset from UCI for the detection of breast cancer. The dataset contains 683 instances and 10 attribute. Three data mining techniques were used and SMO shows the better results as compared to KNN and BF tree. The author used the WEKA tool for the experiment purpose.

U. Karthik Kumar, M.B. Sai Nikhil, and K. Sumangali [6] used the voting ensemble method using three classifiers i.e. NB, SVM, and J48. The author concluded that ensemble method is the best approach for the prediction of Breast cancer. **Uma Ojha, Savita Goel [7]** presented a comparative analysis between four classification and four clustering techniques. The author used the UCI breast cancer dataset and results shows that decision tree and SVM gives better results. Analysis also shows that classification algorithm perform well as compared to clustering techniques.

III. PROPOSED WORK

The process flow Fig.1 shows the proposed model that follows six steps. Both conceptual and empirical approach has been used. The conceptual approach has been used to study the literature surveys, research papers, articles. The empirical approach involved the tools and performance analysis of the breast cancer dataset.

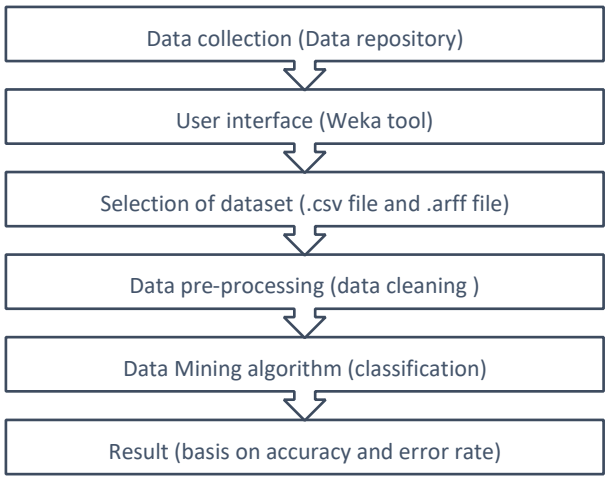

Fig 1. Process flow of proposed model

1. Data Collection – The first step is to collect the raw data. A huge volume of datasets is available on the internet. One of the most famous repositories is the UCI machine learning repository where a large number of datasets are available ex. Healthcare dataset, iris, etc. Breast cancer dataset having attribute 10 collected from the UCI machine learning repository.

2. User Interface – The second step is to use a tool i.e. WEKA.

WEKA is a famous machine learning and data mining tool which is originally developed at the University of Waikato in New Zealand. WEKA contains several supervised and unsupervised learning algorithms like classification, regression, clustering, and association rule. WEKA supports its file formats in .ARFF, .CSV, .json etc. WEKA provides a various application for implementation purpose i.e. Explorer, Experimenter, knowledge flow, etc.

3. Data pre-processing- The next step is data preprocessing. After selecting the dataset, preprocessing is used to convert the raw data into a structured format. Data cleaning, Data transformation are the various steps of data preprocessing. WEKA support data pre-processing filter supervised, and unsupervised.

4. Data mining algorithm- After cleaning the dataset or removing the unwanted data and the missing values from the dataset, the next phase is to apply the data mining techniques. Various classification algorithms are available including Decision tree, Random Forest, Support Vector Machine, etc.

5. **Performance estimation-** the performance estimation is done based on True positive, false Positive, precision, and recall. The correctly classified instance is considered as the accuracy.

IV. METHODOLOGY

We have chosen the breast cancer dataset from the UCI machine learning repository using the WEKA tool. The supervised classification algorithm is applied to the dataset. After applying the classification algorithm we have calculated the accuracy of classifiers. The classifiers which we have used are KNN, RF, and NB. The performance evaluation is based on the correctly classified instance and the incorrectly classified instance. Cross-validation is used for training and testing the dataset.

1. Data Set - UCI machine learning repository provides us with several breast cancer datasets. The dataset contains 286 instance and 10 attributes.

Table 1. Breast cancer recurrence dataset

S.no	Attribute	Type
1.	Age	nominal
2.	Menopause	nominal
3.	Tumor size	nominal
4.	Inv-nodes	nominal
5.	Node caps	nominal
6.	Degree of malignancy	nominal
7.	Breast	nominal
8.	Breast quadrant	nominal
9.	Irradiation	nominal
10.	Class	Recurrence event and non-recurrence event

2. Classification- A Classification algorithm is a supervised learning algorithm that is mainly used to predict categorical values. It is a procedure to predict the class from the given data set. The classification technique aims to predict precisely the target class of objects of which the class label is unknown [8]. Three classifiers that we have used are discussed below.

(i) K-Nearest Neighbors classifier- K-NN classifies the new data based on its nearest available data. It takes new data and classifies the new data into the category of the most similar

data which is already available. The nearest neighbor is determined by calculating the Euclidean Distance and new data can be classified [9]. K-NN also called IBK algorithm in WEKA tool.

(i) Naïve Bayes- Naïve Bayes algorithm is based on the Bayes' Rule and it is depend on the conditional probability. Naïve Bayes is a combination of algorithm where every variable is independent of each other called as class conditional independence [10].

(ii) Random Forest- Random forest is a popular machine learning algorithm that belongs to the supervised learning technique. It combines multiple classifies to solve a complex problem and to improve the performance of the model. It is a type of classifier that contains various decision trees on a various subset of the given dataset and take the average to improve the predictive accuracy of the dataset.

V. EXPERIMENT RESULTS

The Breast Cancer dataset from the UCI machine learning repository is used for analysis. For cleaning of missing values preprocessing of data set is done on WEKA explorer. The Breast cancer dataset is classified by using three classification techniques i.e. NB, k-NN, RF. The value of K is taken 3. For testing purpose, we used cross-validation value 10. Table 2 shows correctly classified instance, incorrectly classified instance, FP, TP, etc. K-NN is the best predictor as compared to other classifiers. Measures for Performance evaluation

$Accuracy=(TP+TN)(TP+TN+FP+FN)$

$Sensitivity=TP(TP+FN)$

$Specificity=TN(TN+FP)$

Table 2: Comparison of classifiers

Classifier	KNN Classifier	Random Forest	Naïve Bayes
Correctly Classified Instances/accuracy	73.7762%	69.5804 %	71.6783 %
In Correctly Classified Instances	26.2238%	30.4196 %	28.3217 %
TPR	0.738	0.696	0.717
FPR	0.552	0.543	0.446
Precision	0.722	0.664	0.704
Recall	0.738	0.696	0.717

VI. CONCLUSION

Table 2, shows the percentage of correctly classified instances of KNN, Random forest, and Naïve Bayes is 73.7762 %, 69.5804 %, 71.6783 %, and incorrectly classified instances are 26.2238%, 30.4196 %, 28.3217 %. The table also shows the KNN has the highest True Positive rate and Random forest has the lowest True Positive rate. From the results, it is observed that K-NN is the best predictor and outperformed as compared to NB and RF in all cases of TP, FP, Precision, and recall. According to the current study, there are some issues like missing values in the datasets. All the algorithms ran successfully on WEKA Explorer. There is other data mining algorithms that can be used for prediction purposes and we can also use the ensemble method for better results. The selection of good algorithms with a good dataset gives better prediction results.

REFERENCES

[1] Nidhi R. Sharma, K. L. Bansal, "Performance of Data Mining Tools: A Case Study Based on Classification Algorithms and Datasets ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 10, October-2015 ISSN: 2277 128X

[2] Shelly Gupta, Dharminder Kumar, and Anand Sharma, "performance analysis of various data mining classification techniques on health care data", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 4, August 2011

[3]S.Joshi, M.K. Nair, "prediction of heart disease using classification based data mining techniques", Smart Innovation, Systems and Technologies 32, 2015 DOI 10.1007/978-81-322-22088_46

[4]Nour A AbouEINadar, Amani A Saad "Towards a Better Model for Predicting Cancer Recurrence in Breast Cancer Patients", Intelligent computing proceedings of the computing conference, 2019, doi.org/10.1007/978-3-030-22871-2_63

[5]Vikas Chaurasia and Saurabh Pal, "A Novel Approach for Breast Cancer Detection Using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering , Vol. 2, Issue 1, January 2014

[6]U. Karthik Kumar, M.B. Sai Nikhil, and K. Sumangali,"Prediction of Breast Cancer using Voting classifier Technique", IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Mateials (ICSTM),2017.

[7]Uma Ojha, Dr. Savita Goel," A study on Prediction of Breast cancer Recurrence using data mining technique", IEEE. 2017 978-1-5090-3519-9/17/$31.00_c

[8]G. Kesavaraj and S. Sukumaran,"A Study on Classification Techniques in Data Mining", IEEE, 2013.

[9]Sarangam Kodati, R. Vivekanandam," Analysis of Heart Disease using in Data Mining Tools Orange and Weka ", Global Journal of Computer Science and Technology, 2018.

[10]M. Narasimha Murty, V. Susheela Devi, "Pattern Recognition: An Algorithmic Approach", Springer Science & Business Media, May 25, 2011.